

"Express Mail" mailing label number EV339771136US

Date of Deposit: December 8, 2003

Our Case No. 9333/361

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE  
APPLICATION FOR UNITED STATES LETTERS PATENT

INVENTORS:                      SHINGO KIUCHI  
                                         NOZOMU SAITO

TITLE:                              SPEECH RECOGNITION  
                                         PERFORMANCE IMPROVEMENT  
                                         METHOD AND SPEECH  
                                         RECOGNITION DEVICE

ATTORNEY:                      JAMES P. NAUGHTON  
                                         Registration No. 30,665  
                                         BRINKS HOFER GILSON & LIONE  
                                         P.O. BOX 10395  
                                         CHICAGO, ILLINOIS 60610  
                                         (312) 321-4200

## SPEECH RECOGNITION PERFORMANCE IMPROVEMENT METHOD AND SPEECH RECOGNITION DEVICE

### BACKGROUND OF THE INVENTION

#### 1. Field of the Invention

**[0001]** The present invention relates to a speech recognition device and a speech recognition performance improvement method. More particularly, the present invention relates to a speech recognition device for improving speech recognition performance in a noisy environment and a speech recognition performance improvement method therefor.

#### 2. Description of the Related Art

**[0002]** Speech recognition devices, by which an operation of vehicle-mounted devices such as audio devices, navigation systems, etc., is performed using speech, have been put into practical use. Fig. 6 is a block diagram of such a speech recognition device. A microphone 1 for entering speech detects speech by a speaker and generates a speech signal. An A/D converter 2 converts the speech signal into digital form. An operation section 3 instructs the starting of speech recognition by operating a switch (not shown). A speech recognition engine 4 recognizes entered speech when the starting of speech recognition is instructed.

**[0003]** An example of the speech recognition engine 4 is disclosed in Japanese Unexamined Patent Application Publication No. 59-61893. In this conventional technology, speech recognition is performed by comparing a feature pattern for each of a series of single syllables in word entered speech with a standard pattern, and by referring to a word dictionary the recognized result is output as a word having a meaning.

**[0004]** In a case where noise is superimposed on speech data that is entered to a speech recognition system, if the speech data is entered to a speech recognition engine by changing the start position of a speech region, such as a portion of a non-speech region which is a start portion of the data being deleted (by changing the length of the non-speech region), there are cases in which the recognized result changes. That is, even in the case of the same produced speech, the correctness of

the recognized result is changed depending on the speech-producing timing (the start position of the speech region).

**[0005]** This phenomenon hardly appears in a case where the magnitude of noise that is superimposed onto speech data, for example, noise inside a vehicle, is sufficiently small with respect to the speech (the S/N ratio is high), but when the magnitude of noise inside a vehicle is large with respect to the speech (the S/N ratio is low), this phenomenon appears conspicuously. The reason why such a phenomenon occurs is that, when the speech recognition engine 4 measures the noise level of the background in a non-speech region \*SIT (Fig. 7) and performs a speech recognition process on the speech data of the speech region SIT, that noise level is used. The non-speech region \*SIT is a region from the time  $t_B$  at which the starting of speech recognition was instructed using a switch to the starting position (speech-producing timing)  $t_{ST}$  of the speech region SIT.

**[0006]** Since this measurement of noise data is a measurement at a region of a short time, even in the case of the noise under the same conditions, measured results vary depending on the measurement position. For this reason, the recognized results vary, with the result that the data may be recognized correctly or incorrectly. For example, as shown in Fig. 7, if the noise level is assumed to be an average level of the non-speech region \*SIT and speech recognition is performed using speech data in the speech region SIT by taking the noise level into consideration, since the noise level is high at the start point of the non-speech region \*SIT in Fig. 7, the shorter the non-speech region \*SIT, that is, the earlier the speech-producing timing  $t_{ST}$ , the higher the average level; and the longer (the later) the speech-producing timing  $t_{ST}$ , the lower the average level becomes. In the manner described above, the level of the noise to be measured varies depending on the speech-producing timing  $t_{ST}$ , and as a result, the correctness of the recognized result changes.

**[0007]** The above phenomenon shows a situation in which, even in an environment where a certain degree of S/N is ensured, incorrect recognition occurs due to the timing of the speech production. When viewed from the user side, the recognition performance is decreased, which causes a problem.

[0008] In the conventional technology, including the technology of Japanese Unexamined Patent Application Publication No. 59-61893, improvement of the recognition rate is sought by exclusively increasing the recognition accuracy of the speech recognition engine, but there are limits.

## SUMMARY OF THE INVENTION

[0009] Accordingly, an object of the present invention is to improve the speech recognition performance without changing the speech recognition engine.

[0010] In the speech recognition device, a plurality of pieces of speech data whose start positions of the non-speech regions differ are generated from speech data for which speech recognition is to be performed. Speech recognition is performed by using each of the pieces of speech data, and the most numerous recognized result from among a plurality of obtained recognized results is provided as an output. As a result of the above, since the start position of the non-speech region is shifted, although there may happen to be speech data which is recognized incorrectly, if a large number of pieces of speech data are recognized and the numbers thereof are compared, the number of cases in which the speech data is recognized correctly becomes the most numerous. Therefore, by providing the result which is recognized most often, the recognition performance can be improved without changing the recognition engine.

[0011] In order to generate a plurality of pieces of speech data whose start positions of non-speech regions differ, the start position of the non-speech region is shifted in sequence from the start position of the speech region to a preceding position by a predetermined time. That is, the input speech signal is A/D-converted at a predetermined sampling speed, and this speech signal is stored in the buffer in the order of sampling. Then, a plurality of pieces of speech data whose start positions of non-speech regions differ is generated by changing the position at which reading from the speech buffer starts.

[0012] The speech recognition process of each of the pieces of speech data may be performed by one speech recognition engine, but it takes time. In order to shorten the processing time, a speech recognition engine is provided so as to

correspond to each of the pieces of speech data whose start positions of the non-speech regions differ, and the most numerous recognized result from among the recognized results of each speech recognition engine is supplied as an output.

[0013] As described above, according to the speech recognition device of the present invention, the speech recognition performance can be improved without changing the speech recognition engine.

[0014] As described above, according to the present invention, since a plurality of pieces of speech data whose start positions of the non-speech region differ is generated from speech data for which speech recognition is to be performed, speech recognition is performed by using each of the pieces of speech data, and a plurality of recognized results are obtained. Thus, the recognition performance can be improved without changing the recognition engine.

[0015] Furthermore, according to the present invention, by providing a speech recognition engine in such a manner as to correspond to each of the plurality of pieces of speech data, a speech recognition result can be obtained at a high speed, and moreover, recognition performance can be improved.

[0016] Furthermore, according to the present invention, the phenomenon of incorrect recognition due to the speech-producing timing when a recognition engine is used in an environment in which a certain degree of S/N or higher (2 to 3 dB or higher) is ensured can be eliminated. When viewed from the user side, this is the same as having the same advantages as the recognition performance being improved in a noisy environment, and thus, the present invention is useful.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Fig. 1 is a block diagram Fig. 1 is a block diagram of a speech recognition device according to a first embodiment of the present invention;

[0018] Fig. 2 is an illustration of a speech data generation section;

[0019] Fig. 3 is an illustration of a speech signal;

[0020] Fig. 4 is a block diagram of a totaling section and a comparison section;

[0021] Fig. 5 is a block diagram of a speech recognition device according to a second embodiment of the present invention;

[0022] Fig. 6 is a block diagram of a conventional speech recognition device;  
and

[0023] Fig. 7 is an illustration of a speech region and a non-speech region.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

### (A.) First Embodiment

[0024] Fig. 1 is a block diagram Fig. 1 is a block diagram of a speech recognition device according to a first embodiment of the present invention.

Fig. 2 is an illustration of a speech data generation section. When the starting of speech recognition is instructed by operating a switch, a microphone 11 for entering speech detects speech by a speaker and generates a speech signal. An A/D converter 12 performs A/D conversion on the speech signal (see Fig. 3) at a predetermined sampling speed. A speech buffer 13 stores the A/D-converted speech data in the order of sampling. The speech data is generated in such a manner that, as shown in Fig. 3, a speech signal (noise) of the non-speech region \*SIT and a speech signal of the speech region SIT are sampled chronologically, and these are assigned with a number 1 to n in sequence and are stored in sequence in the speech buffer 13, as shown in Fig. 2. The data with an earlier number is data of the non-speech region \*SIT, and the data with a later number is data of the speech region SIT.

[0025] A speech data generation section 14 generates a plurality of pieces of speech data DT1, DT2, DT3, . . . . . whose starting positions of the non-speech regions differ, by shifting the position at which reading from the speech buffer 13 starts, and stores the plurality of pieces of speech data in a speech data storage section 15. The shift point of the reading start position is from the start position  $t_{ST}$  of the speech region SIT back to a position  $t_C$  preceding  $t_{ST}$  by the predetermined time T, as shown in Fig. 3. A speech recognition engine 17 of a recognition processing section 16 identifies the start position  $t_{ST}$  of the speech region SIT, determines the final shift position  $t_C$  by using the start position  $t_{ST}$ , and provides it as speech start-point detection information SST. The speech start-point detection information SST is obtained each time recognition processing is

performed on one piece of speech data. The speech start-point detection information when recognition processing is performed on the first speech data may be used, and information such that speech start-point detection information for several pieces of data from the start is averaged may be used.

**[0026]** Referring to Fig. 2, a pointer control section 21 of the speech data generation section 14 receives the speech start-point detection information SST (= the final shift position  $t_c$  of the reading start), which is supplied from the speech recognition engine 17. The pointer control section 21 shifts the reading position (pointer) in sequence, and supplies the reading position to a data reading section 22. The data reading section 22 reads, from the speech buffer 13, sampling data on the basis of the position indicated by the specified pointer, and stores the sampling data in the speech data storage section 15. When the reading of one piece of speech data is completed, the pointer control section 21 shifts the reading start position by one piece of sampling data in order to shift the reading position (pointer) in sequence, and supplies the reading position to the data reading section 22. The data reading section 22 reads, from the speech buffer 13, sampling data on the basis of the input position indicated by the pointer, and stores the sampling data in the speech data storage section 15. Hereafter, each time the reading of the speech data is completed, the reading position (pointer) is shifted to read the speech data. When the reading position (pointer) becomes equal to the final shift position  $t_c$ , the process of generating the speech data is completed.

**[0027]** Concurrently with the above, the speech recognition engine 17 of the recognition processing section 16 performs a speech recognition process by using the first speech data DT1, detects the start position  $t_{ST}$  of the speech region SIT, and generates the speech start-point detection information SST. Then, the recognized result (recognized result 1) is stored in a recognized result storage section 18.

**[0028]** Next, the speech recognition engine 17 performs a speech recognition process by using the second piece of speech data DT2 and stores the recognized result 2 in the recognized result storage section 18. Hereafter, in a similar manner,

the recognized results 1 to k of all the speech data DT1 to DTk are stored in the recognized result storage section 18.

**[0029]** When the recognition of all the speech data DT1 to DTk is completed, a totaling/comparison section 19 provides, as the final result, the most numerous recognized results from among a plurality of the obtained recognized results.

Fig. 4 is a block diagram of the totaling/comparison section 19, which has a totaling section 31 and a compared result output section 32. The totaling section 31 totals the number for each compared result. In Fig. 4, the numbers of compared results A, B, and C are p, q, and r, respectively. The compared result output section 32 provides, as the final recognized result, the recognized result corresponding to the maximum value from among p, q, and r.

**[0030]** As described above, according to the first embodiment, since the start position of the non-speech region is shifted, although there may happen to be speech data which is recognized incorrectly due to the influence of noise, the number of cases which the speech recognition engine, that correctly recognizes speech data when noise is not present, recognizes correctly becomes the most numerous if a large number of pieces of speech data are recognized and the numbers thereof are compared. Therefore, by providing the result which is recognized most often, the recognition performance can be improved without changing the recognition engine.

#### (B) Second Embodiment

**[0031]** Fig. 5 is a block diagram of a speech recognition device according to a second embodiment of the present invention. Components in Fig. 5 which are the same as those in Fig. 1 in the first embodiment are designated with the same reference numerals. The points differing from those of the first embodiment are as follows:

**[0032]** (1) The speech data storage section 15 for storing the speech data DT1 to DTk is omitted,

**[0033]** (2) k recognition engines  $17_1$  to  $17_k$  are provided so as to correspond to k pieces of speech data received from the speech data generation section 14,



**[0034]** (3) The recognition engines  $17_1$  to  $17_k$  perform a speech recognition process on  $k$  pieces of speech data and provide each of the recognition results A, B, C . . . to the totaling/comparison section 19, and

**[0035]** (4) The totaling/comparison section 19 supplies, as the final recognized result, the most numerous recognized result from among the recognized results of the recognition engines  $17_1$  to  $17_k$ .

**[0036]** As a result of providing  $k$  speech recognition engines in this manner, a speech recognition result can be obtained at a high speed, and moreover, recognition performance can be improved.